Effervescence autour des corpus

Damon Mayaffre

▶ **To cite this version:**

Damon Mayaffre. Effervescence autour des corpus. Michel Ballard et Carmen Pineira (dir.). Corpus en linguistique et en traductologie, Artois Presses Université, pp.61-71, 2007. hal-00912006

HAL Id: hal-00912006
https://hal.science/hal-00912006

Submitted on 1 Dec 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Michel Ballard et
Carmen Pineira-Tresmontant

LES CORPUS EN LINGUISTIQUE ET EN TRADUCTOLOGIE

Artois
Presses
Université
Traductologie

EFFERVESCENCE AUTOUR DES CORPUS

Damon Mayaffre
CNRS - UMR 6039 « Bases, Corpus et Langage » (Nice)

Introduction

La réflexion sur les corpus semble devenue incontournable aujourd'hui en linguistique. L'année 2004-2005 a connu, outre les journées « Corpus en linguistique et en traductologie » d'Arras plusieurs colloques, réunions, tables rondes portant sur le sujet[1]. Plusieurs publications sont aussi venues enrichir une bibliographie française ouverte par le livre de (Habert *et alii* 1997). L'ouvrage collectif sous la direction de (Williams 2005-a), *La Linguistique de corpus*, présente même par son introduction et certaines de ses contributions un premier recul épistémologique important pour un domaine plus seulement en pleine expansion mais en pleine structuration.

Cette effervescence scientifique trouve une reconnaissance institutionnelle, et, de manière significative, le CNRS met au concours pour la campagne 2006, un poste CR1 coloré « linguistique de corpus ». Une revue aussi prend son essor dont le nom lui-même fait programme : *Corpus* ; riche aujourd'hui de 4 numéros, elle se consacre, précisément, depuis 2001, « à la linguistique

[1] On citera : Journée scientifique « Corpus de Sciences sociales : établissement, numérisation, analyses sémantiques », INALCO, Paris, 8 juin 2004 ; École d'été du CNRS « Linguistique de Corpus », Université de Caen, 14-19 juin 2004 ; COL'DOC'2005, « Recueil des données en Sciences du langage et constitution de corpus : données, méthodologie, outillage », Paris, 16-17 juin 2005 ; JETOU'2005 « Rôle et place des corpus en linguistique », Toulouse, 1-2 juillet 2005 ; 1er Journées « Corpus en linguistique et en traductologie », Arras , 28-29 octobre 2005 ; Journées « Corpora et questionnements littéraires » organisées par le Sit@t et Modyco, Paris, 15-16 novembre 2005.

de corpus... envisagée sous tous ses aspects : théoriques, épistémologiques, méthodologiques. » (Mellet 2001 : 5)[2].

À cela ajoutons enfin que cette dynamique n'est pas uniquement française. Depuis quelques années déjà, dans le monde anglo-saxon, fleurissent des publications se revendiquant de pratiques linguistiques explicitement fondées sur les corpus (Biber, Conrad & Reppen 1998 ; Tognini-Bonelli 2001 ; Aijmer and Altenberg (ed.) 2002).

Longtemps mal considérés par une partie de la discipline, toujours sous le coup de la critique chomskyenne, les corpus tiennent donc aujourd'hui une forme de revanche. Le propos de cette communication est d'essayer de la comprendre mais aussi de la modérer, car si le corpus et la linguistique de corpus apparaissent actuellement comme un sésame ouvrant les portes sur les questions les plus profondes de la discipline (la question de l'objet de la linguistique, la question de la méthode pour traiter cet objet, la question épistémologique de l'empirie et de la théorie ou de l'induction et de la déduction, comme nous avons essayé de le montrer dans (Mayaffre 2005-a et 2005-b), ils ne sauraient être considérés, sans précaution, comme une panacée susceptible de résoudre tous les problèmes épistémologiques de nos pratiques.

1. Le corpus parasite l'image idéale du système

Quand bien même le corpus serait-il devenu un lieu incontournable pour le linguiste, il reste un lieu intrinsèquement subversif. En renonçant à l'introspection et en travaillant sur corpus, le théoricien renonce sans doute à la pureté du système qu'il entendait décrire. Si la linguistique est l'étude de la langue en elle-même, pour elle-même, selon les propos que Bailly prête à Saussure, alors le corpus est problématique car aucun système complexe ne peut se laisser enfermer dans le champ clos, nécessairement partiel, toujours arbitraire d'un corpus de textes ou de données.

Restrictifs par leur clôture d'un côté, les grands corpus de textes apportent, aussi, de l'autre côté, des exceptions à l'impossible modèle idéal que la langue est censée représenter. À la Loi linguistique, ils opposent des jurisprudences. À l'unicité, ils opposent l'hétérogénéité. À la synchronie, des variations multiples (variations temporelles, sociales, individuelles).

Bref, la langue est pure, mais le corpus la confond par d'infinies situations de communication, pratiques discursives diverses, genres de textes démultipliés, langues de spécialité, etc.

---

[2] *Corpus* est publiée par l'UMR 6039 « Bases, Corpus et Langage » (CNRS / Université de Nice-Sophia Antipolis). Un an après leur diffusion papier, les numéros parus sont consultables en ligne à l'adresse : http://revel.unice.fr/corpus/.

Finalement, le corpus non seulement ne permet pas d'accéder au système, mais nous empêche définitivement d'y accéder par l'image immanente, brouillée, interprétée pourrait-on dire, qu'il propose de la langue. En révélant le système sous un jour nécessairement particulier, le corpus déjà le trahit en faisant la part belle à l'individuel, à l'exception, à la déviance.

Sans doute est-ce pour cette raison que le corpus fut boudé et la linguistique de corpus niée. Chomsky lui-même, fidèle à sa position mais dans un combat déjà d'arrière garde, déclarait encore en 1999, « corpus linguistics does not exist » (Chomsky, entretien avec Baas Aarts, cité par Rastier 2005 : 40). De fait, à moins de travailler sur des corpus sur mesure, faits d'exemples forgés, non pas trouvés mais controuvés, aucun corpus ne peut offrir une image non altérée, encore moins exhaustive du système. C'est en prenant acte de cette réalité, d'ailleurs, que les syntacticiens soit ignorent les corpus, soit en construisent de toute pièce. Dans ce second cas, ils peuvent prétendre eux aussi travailler sur corpus, mais l'on comprend combien ici la notion ne recouvre pas la même valeur que celle des inventeurs de la « linguistique de corpus » au sens des auteurs susnommés Habert *et alii* (1997), Biber, Conrad & Reppen (1998) ; Tognini-Bonelli (2001) ; Rastier (2005) ; etc.). Car de manière définitive, Williams (2005-b : 13) affirme : « La linguistique de corpus est un domaine qui s'intéresse aux textes, aux textes réels », requalifiant ainsi, au passage, l'objet même d'une linguistique adulte qui ne saurait s'arrêter aux limites du signe ou de la phrase[3].

2. Les corpus : une double attestation des données

Travailler sur corpus, c'est exiger de travailler sur des données langagières attestées. Cette affirmation évidente se complique si l'on veut bien considérer qu'elle s'entend dans un double sens.

D'abord, nous avons affaire à des recueils de données qui ont été effectivement émises par des locuteurs authentiques. Dès lors, il ne peut s'agir que de textes ou de discours[4] puisque nous nous exprimons non pas avec des

[3] Nous touchons là, sans doute, au postulat fondamental de la linguistique de corpus que nous avons développé ailleurs (Mayaffre 2005-a et –b). Pour de plus en plus de chercheurs, le texte – sinon le corpus textuel – devient l'objet ultime du linguiste. Au départ : Hjelmslev 1943 (1968-1971) et Bakhtine 1952-1953 (1984). Ensuite dans des perspectives différentes : (Halliday et Hasan 1976),(Van Dijk 1984), [(Combettes 1983), (Bronckart 1997), (Rastier 1989 et 2001) ; (Amossy 2002) ; (Adam 1990 et 1999) ; etc. On se reportera nécessairement à ce dernier pour la bibliographie commentée de cette linguistique du texte : [Adam 1999 : *Introduction* : 5-20 et *Chapitre I : Pour une linguistique des grandes unités* : 21-42).

[4] Nous n'entrerons pas ici sur le distinguo texte/discours en renvoyant à une bibliographie foisonnante sur le sujet et à l'ouvrage le plus récent sur la question : (Adam et Heidmann 2005).

mots ou des phrases mais avec des séquences motivées de longueur et de structures indéterminées que l'on appellera texte ou discours. Ces textes sont donc attestés dans le sens où ils ont été réellement produits par un locuteur de chair et d'os, historiquement, culturellement, psychologiquement situé, et reçus par un auditeur tout aussi authentique[5]. Partant, précisons-le, la linguistique de corpus s'inscrit sans doute dans une perspective trans- ou inter- disciplinaire au sens de Darbellay (2005)[6]. Car en admettant que les conditions historiques de production-réception des textes en font partie intégrante, l'analyse peine à se contenter de considérations uniquement linguistico-linguistiques.

En tous cas, on le comprend, c'est bien l'usage qui nous intéresse plus que le système, le réel du langage plus que les possibles théoriques de la langue, le choix linguistique plutôt que le prédictible, voire la distorsion (si l'on admet que la langue est un système clos et invariable de règles) plutôt que les jugements (souvent contestables) de grammaticalité. Selon les mots programmatiques de Marie-Paule Jacques (2005 : 27) :

> La linguistique de corpus s'inscrit dans une certaine conception de la langue et des objectifs même de la linguistique, elle prend sens dès lors que l'on pense la langue non comme UN système désincarné et abstrait mais comme un ensemble vivant, peut-être multiforme, où la description de la variation et de la multiplicité des usages peuvent être aussi fructueux pour la découverte des règles que les raisonnements sur les possibles et les impossibles.

Ensuite, l'attestation des données se comprend à un deuxième niveau : en corpus, nous avons affaire à des données sélectionnées, saisies, organisées par l'analyste. Le corpus est en lui-même attestation (attestation secondaire donc, qui vient en plus de celle, initiale, produite par le locuteur en train de parler ou d'écrire). Le corpus est matérialité (voir *infra*). Il matérialise, pour l'analyste, un état de langue. La réalité du corpus (son attestation donc) est la condition de la manipulation de l'objet scientifique sans laquelle aucune étude élaborée n'est possible.

---

[5] À vrai dire affirmer qu'un « texte doit être attesté » est tautologique. Si les mots peuvent être appréhendés, *in abstracto*, dans un dictionnaire, et les phrases (plus ou moins fabriquées) dans des exempliers, un texte n'est compréhensible que dans son contexte : il est nécessairement authentique et attesté, et son attestation fait partie intégrante de sa définition, de sa compréhension, de son interprétation.

[6] L'auteur d'ailleurs, de manière significative, décrit la posture inter- et trans-disciplinaire en SHS *via* l'analyse des textes et des discours. C'est par le biais de la « complexité des textes », de «l'intertextualité et la transtextualité » qu'il dénonce le cloisonnement monodisciplinaire.

Par là et plus profondément, le corpus, dans sa manière, arbitraire mais consciente, d'attester ensemble tels et tels textes (et d'exclure tels ou tels autres), est le lieu où se construit le sens et s'organise le savoir. C'est sous la plume de François Rastier que nous trouvons les propos les plus avancés sur le sujet : « Tout texte placé dans un corpus en reçoit des déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent » (Rastier 2001 : 92). Ou, pour resituer la réflexion dans le panorama général de la théorie rastirienne du *local* et du *global* : « La détermination du local par le global s'exerce en somme de deux façons, par l'incidence du texte sur ses parties, par l'incidence du corpus sur le texte. » (Rastier 2001 : 109). Et l'auteur de conclure dans son article « Enjeux épistémologiques de la linguistique de corpus » : « Le texte est pour une linguistique évoluée l'unité minimale, et le corpus l'ensemble dans lequel cette unité prend son sens » (Rastier 2005 : 31).

Loin d'être un recueil mort, le corpus est un objet vivant, producteur de sens qui permet, autant que possible, d'objectiver l'intertexte nécessaire à l'interprétation des textes (voir nécessairement *Cahiers de praxématique* (1999). « Sémantique de l'intertexte »). C'est ainsi, nous semble-t-il, que la linguistique de corpus entretien des rapports privilégiés avec l'herméneutique : elle vise à comprendre la production du sens, en pensant l'organisation des parcours interprétatifs. Dans ce parcours, le choix, l'organisation, la structuration, l'outillage du corpus et de ses composantes jouent un rôle déterminant : la linguistique de corpus, en théorisant le corpus, se propose, précisément, de penser cette organisation pour contrôler l'interprétation.

Notons pour finir que par cette double attestation des textes qu'il propose (attestation primaire : celle du locuteur qui les a produits ; attestation secondaire : celle de l'analyste qui les a rassemblés), le corpus peut être considéré aussi bien comme un lieu théorique qu'empirique. Scheer (2004), renversant une idée reçue, a récemment insisté sur la nature théorique des corpus. Toujours construits (et non pas donnés), sur la base d'hypothèses de travail explicites, le corpus suppose en amont une intuition linguistique forte. Par là, il est, pour l'auteur, l'observatoire construit de phénomènes linguistiques pressentis. Il est un outil, façonné, explicitement à dessein, pour mettre à l'épreuve un savoir ou une supposition théorique. Dans la terminologie de Tognini-Bonelli (2001) l'approche que Scheer propose est *corpus-based* : l'on renonce certes à la pure introspection mais le corpus est avant tout considéré comme une masse documentaire qui sert d'appui à l'analyse pour confirmer ou infirmer des hypothèses.

Le propos de la linguistique de corpus *stricto sensu* semble assez différent, et nous insisterons plus sur la dimension empirique de l'objet, sur sa fonction heuristique et non validante ou illustrative, sur une démarche plus *bottom up* que *top-down*, plus inductive que déductive. L'approche en effet est *corpus-*

driven (Tognini-Bonelli 2001). Le corpus mène l'analyse. Il est l'objet même de la recherche pour la simple raison qu'il est une matrice du sens plus qu'un réceptacle. Il ne révèle pas un monde déjà-là, mais en crée de nouveaux toujours renouvelés. Les corpus sont, en soi, dignes d'intérêt car ils sont pour nous, en eux-mêmes, une quête vers le sens qu'ils construisent.

3. Corpus, sa matérialité, son instrumentation

Mais revenons un instant, pour conclure, sur l'essentiel. Dans les sciences, l'objet tel qu'on peut l'appréhender et l'outil qui nous sert à le disséquer entretiennent des rapports intimes pour circonscrire le champ et la méthode. Dans ce cadre, la linguistique de corpus et la reconsidération des corpus en linguistique nous semblent déterminées par le nouveau support – support électronique comme on le sait – de l'objet linguistique (du simple fichier.txt aux documents numériques structurés, enrichis, balisés, en passant, par exemple, par des bases de données) et l'outillage informatique qui permet de le manipuler (éditeurs XML, concordanciers, lemmatiseurs, analyseurs syntaxiques, outils lexicométriques, etc.). Et ce n'est évidemment pas un hasard si « la linguistique de corpus est une discipline… qui a vraiment pris son essor avec l'arrivée sur le marché d'ordinateurs personnels » (Williams 2005-b : 13).

En matière de corpus, l'ordinateur ne signifie pas virtualisation mais concrétion. Aujourd'hui un corpus n'est pas une notion floue aux contours à géométrie variable. C'est d'abord une saisie – un acte volontaire et effectif de saisie. Sur cette saisie – et uniquement sur celle-ci –, un certain nombre d'algorithmes vont s'appliquer avec l'implacable froideur du système binaire.

Dans le champ de la lexicométrie par exemple, le corpus ne saurait être insaisissable mais est nécessairement défini, entré en machine et pour le besoin du traitement statistique clôturé. Dès lors, le traitement des données sera *exhaustif* et *systématique*. Le traitement quantitatif permet une description objective et des comparaisons formelles entre parties du corpus, le retour contrôlé au contexte grâce aux concordanciers et aux vertus de l'hypertextualité permettent de s'engager vers une herméneutique numérique (Mayaffre 2002-b ; Viprey 2005).

Dès 1997, Habert *et alii* (1997 : 7) ouvraient leur ouvrage, *Les Linguistiques de corpus*, sur la révolution que représentaient la grande taille, l'accessibilité et l'enrichissement des corpus électroniques. Bien sûr, neuf ans après, l'on peut toujours regretter l'absence pour le français d'un corpus de référence à l'égal du Brown corpus pour les USA ou du LOB corpus pour le Royaume-Uni. Mais le foisonnement des corpus et le développement incessant d'outils pour les exploiter permettent plus que jamais d'imaginer que la description et la compréhension des faits linguistiques passeront dorénavant par la confrontation du linguiste avec des données attestées.

Conclusion

Autour des corpus, la problématique s'est aujourd'hui déplacée. Actuellement, la question posée concerne moins leur nécessité que leur nature et que leur fonction. Peu nombreux, même du côté des génératives, sont ceux qui aujourd'hui méprisent les corpus. D'une certaine manière, aujourd'hui, tout le monde fait de la linguistique sur corpus, si l'on en croit par exemple la diversité des praticiens rassemblés lors des dernières réunions sur le sujet (phonologues, sémanticiens, phraséologues, traductologues, etc.).

Le clivage se fait donc à un autre niveau.

Pour certains, un corpus rassemble nécessairement des productions langagières authentiques c'est-à-dire des textes. Pour d'autres, les recueils de mots ou de phrases sont tout aussi recevables. En d'autres termes, la question est de savoir si une base de données constitue un corpus, si la saisie d'un dictionnaire par exemple peut ouvrir sur une linguistique de corpus pleine et entière.

Ces questions renvoient en réalité à une interrogation plus importante sur la fonction des corpus dans le procès de nos recherches. Les corpus sont-ils là seulement pour valider une hypothèse ? Sont-ils des champs d'expérimentation ? Sont-ils avant tout une ressource documentaire ?

Ou au contraire, le corpus constitue-t-il, en soi, un objet digne pour le linguiste, dont la principale vertu est de nous interroger sur des phénomènes langagiers à découvrir ?

RÉFÉRENCES BIBLIOGRAPHIQUES

ADAM, Jean-Michel, *Éléments de linguistique textuelle*. Bruxelles : Mardaga, 1990.
Linguistique textuelle. Des genres de discours aux textes. Paris : Nathan, 1999.

ADAM, Jean-Michel, et HEIDMANN, U, (éd.), *Sciences du texte et analyse de discours. Enjeux d'une interdisciplinité,* Genève, Slatkine, 2005.

AIJMER, Karin et ALTENBERG, B, (éd.), « Advances in Corpus » dans *Corpus Linguistics,* Amsterdam, Rodopi, 2002.

AMOSSY, Ruth, (éd.), *Pragmatique et analyse des textes,* Tel-Aviv, Presses de l'Université de Tel-Aviv, 2002.

BIBER, Douglas, *Variation accross speech and writing,* Cambridge, Cambridge University Press, 1988.
Dimensions of Register Variation: A Cross-linguistic Comparison, Cambridge, Cambridge University Press, 1995.

BIBER, Douglas, CONRAD, S et REPPEN, R, *Corpus linguistics. Investigating language, Structure and Use,* Cambridge, Cambridge University Press, 1998.

BOMMIER-PINCEMIN, Bénédicte, *Diffusion ciblée automatique d'informations: conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents,* Thèse de doctorat, Paris IV, 1999-a.
« Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative » dans A. Condamines *et al* (éd.), *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, Cargèse, Actes de l'atelier thématique TALN, 1999-b, p. 26-36.

BOUQUET, Simon, *Introduction à la lecture de Saussure,* Paris, Payot, 1997.
Après un siècle, les manuscrits de Saussure reviennent bouleverser la linguistique. Texto ! 2005. <http://www.revue-texto.net/Saussure/Sur-Saussure/Bouquet_Apres.html>.

BRONCKART, Jean-Paul, *Activité langagière, textes et discours,* Lausanne-Paris, Delachaux et Niestlé, 1997.

Cahiers de praxématique « Sémantique de l'intertexte », 33, 1999.

COMBETTES, Bernard, *Pour une grammaire textuelle,* Bruxelles, De Boeck-Duculot, 1983.

CORPUS, « Corpus et recherches linguistiques », 1, coordonné par S. Mellet, 2002, 175 pages.

CORPUS, « Corpus en phonologie », 3, coordonné par T. Scheer, 2004, 516 pages.

CORPUS, « Corpus politiques : objet, méthode et contenu », 4, coordonné par D. Mayaffre, 2005, 221 pages.

DALBERA, Jean-Philippe, « Le corpus entre données, analyse et théorie », *Corpus*, 1, 2002, p. 89-105.

DARBELLAY, Frédéric, *Interdisciplinarité et transdisciplinarité en analyse des discours. Complexité des textes, intertextualité et transtextualité,* Genève, Slatkine, 2005.

FABRE, Cécile, HABERT, B. et ISSAC, F, *De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques,* Paris, InterEditions/Masson, 1998.

HABERT Benoît, NAZARENKO, A et SALEM A, *Les Linguistiques de corpus*, Paris, Colin, 1997.

HJELMSLEV, Louis, *Prolégomènes à une théorie du langage,* Paris, Minuit [1943] 1968-1971.

HALLIDAY, Mak et HASAN, R, *Cohesion in English,* Londres, Longman, 1976.

JACQUES, Marie-Paule, « Pourquoi une linguistique de corpus » dans Williams, G (éd.), *La Linguistique de corpus,* Rennes, PUR, 2005, p. 21-30.

MAYAFFRE, Damon, « Les corpus réflexifs : entre architextualité et intertextualité », *Corpus*, 1, 2002-a, p. 51-70.

« L'Herméneutique numérique », *L'Astrolabe. Recherche littéraire et Informatique*, (http://www.uottawa.ca/academic/arts/astrolabe/), 2002-b.

« Les corpus politiques : objet, méthode et contenu. Introduction », *Corpus,* 4, 2005-a, p. 5-19.

« Rôle et place des corpus en linguistique. Réflexions introductives » dans Actes des JETOU 2005 (sous presse), [En ligne sur *Texto !* (http://www.revue-texto.net/)]

MELLET, Sylvie, « Corpus et recherches linguistiques : introduction », *Corpus*, 1, 2002, p. 5-13.

RASTIER, François, *Arts et sciences du texte.* Paris, Puf, 2001.

« Enjeux épistémologiques de la linguistique de corpus » dans G. Williams (éd.), *La Linguistique de corpus,* Rennes, Pur, 2005-a, p. 31-45. [En ligne sur *Texto !* (http://www.revue-texto.net/Inedits /Rastier/ Rastier_Enjeux.html)]

« Discours et texte » (première partie), 2005-b. *Texto !* (http://www.revue-texto.net/Reperes/Themes/Rastier_Discours.html).

SCHEER, Tobias, « Le corpus heuristique : un outil qui montre mais ne démontre pas », *Corpus*, 3, 2004, p. 153-193.

TOGNINI-BONELLI, Elena, *Corpus Linguistics at Work,* Amsterdam, John Benjamin's Publishing, 2001.

VIPREY, Jean-Marie, « Philologie numérique et herméneutique intégrative » dans Adam, J-M et Heidmann, U. (éd.), *Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarit,* Genève, Slatkine, 2005, p. 51-68.

VAN DIJK, Teun, « Texte » dans Beaumarchais *et al* (éd.), *Dictionnaire des littératures de langue française*, Paris, Bordas, 1984.

WILLIAMS, Geoffrey., (éd.) *La Linguistique de corpus,* Rennes, PUR, 2005-a.

« Introduction » dans Williams G (éd.), *La Linguistique de corpus,* Rennes, PUR, 2005-b, p. 13-18.